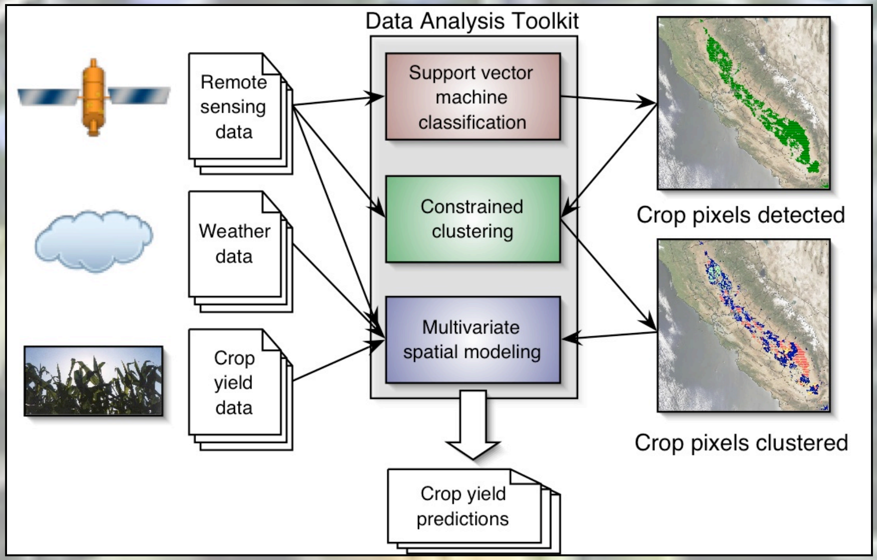
HARVIST: A System for Agricultural and Weather Studies Using Advanced Statistical Methods

Kiri L. Wagstaff

Jet Propulsion Laboratory

Co-Is: Dominic Mazzoni (JPL) and Stephan Sain (Univ. of CO, Denver)

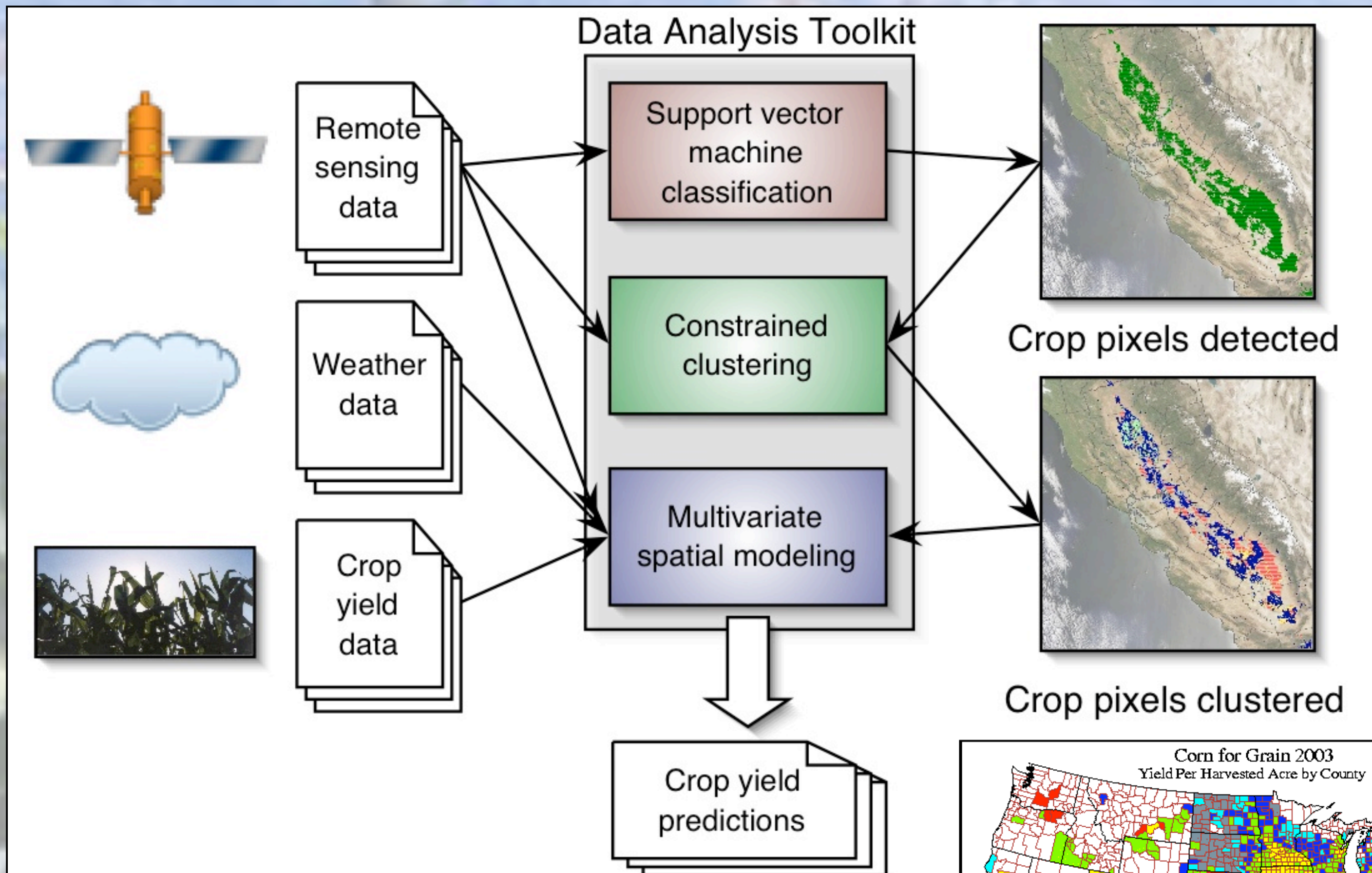
June 28, 2005

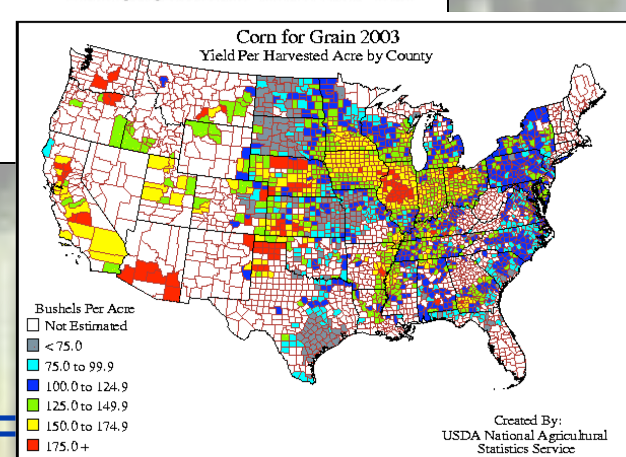


Outline

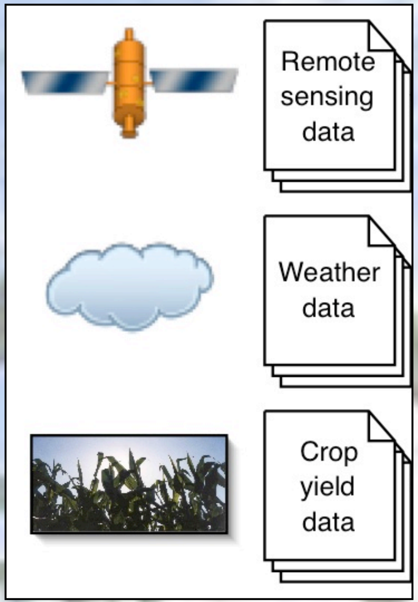
- Project overview
- Preliminary study: corn yield prediction (12 counties)
  – Key finding: integrating data from multiple sources increases prediction accuracy
- HARVIST demonstration
- Accomplishments
  1. Classifier method (SVM) efficiency improvements
  2. Integration of classification and clustering together
     - Example: only cluster regions known to contain vegetation
- Future plans

HARVIST Overview





Key Ideas of the Project

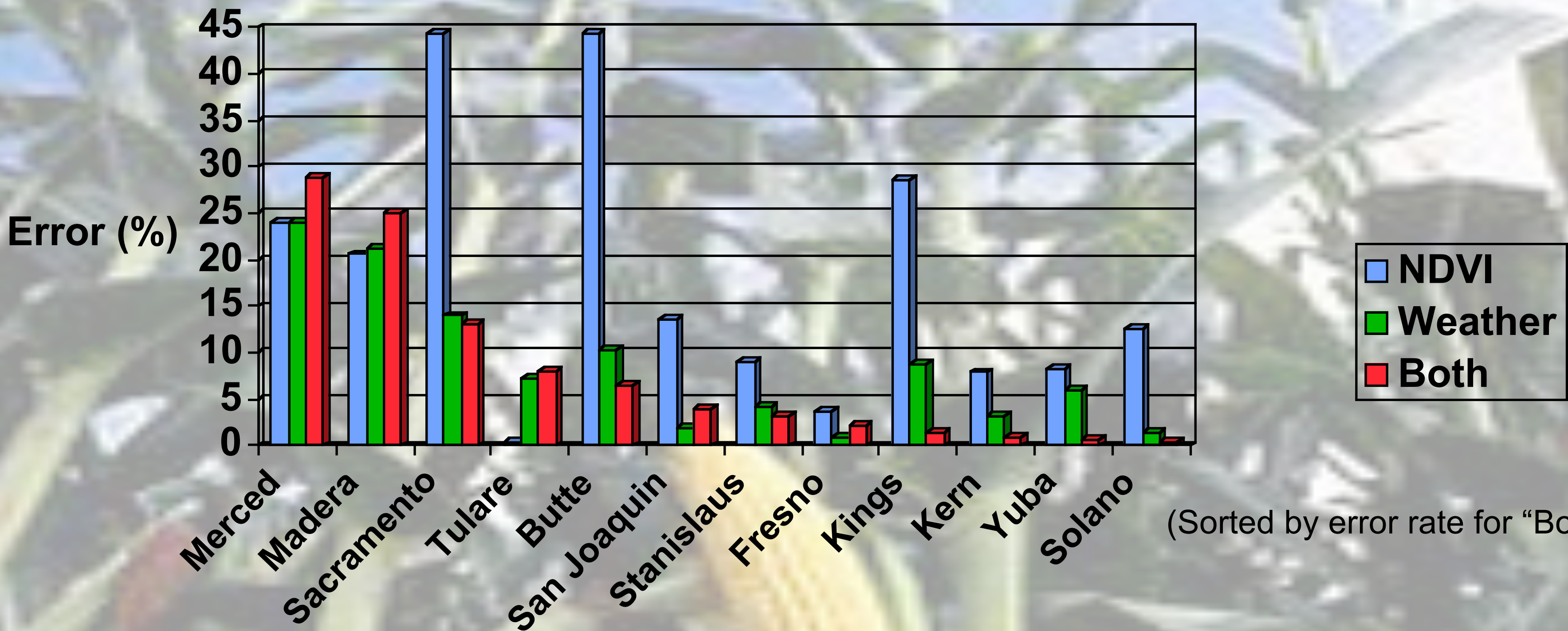


- Combine data from **multiple, diverse sources**:
  - Satellite imagery (LandSat, MODIS, MISR)
  - Weather stations (NCDC)
  - Historical crop yields (USDA)
  - Land cover types (USGS)
  - All at different spatial and temporal resolutions
- Optimize **machine learning techniques for image data**
  - Exploit spatial dependencies to improve efficiency
- **Analyze connections** between weather and agriculture
  - Learn relationships between variables from multiple different data sources
  - Our study: Predict crop yield for different weather conditions

MODIS: MODerate resolution Imaging Spectroradiometer;
MISR: Multi-angle Imaging SpectroRadiometer

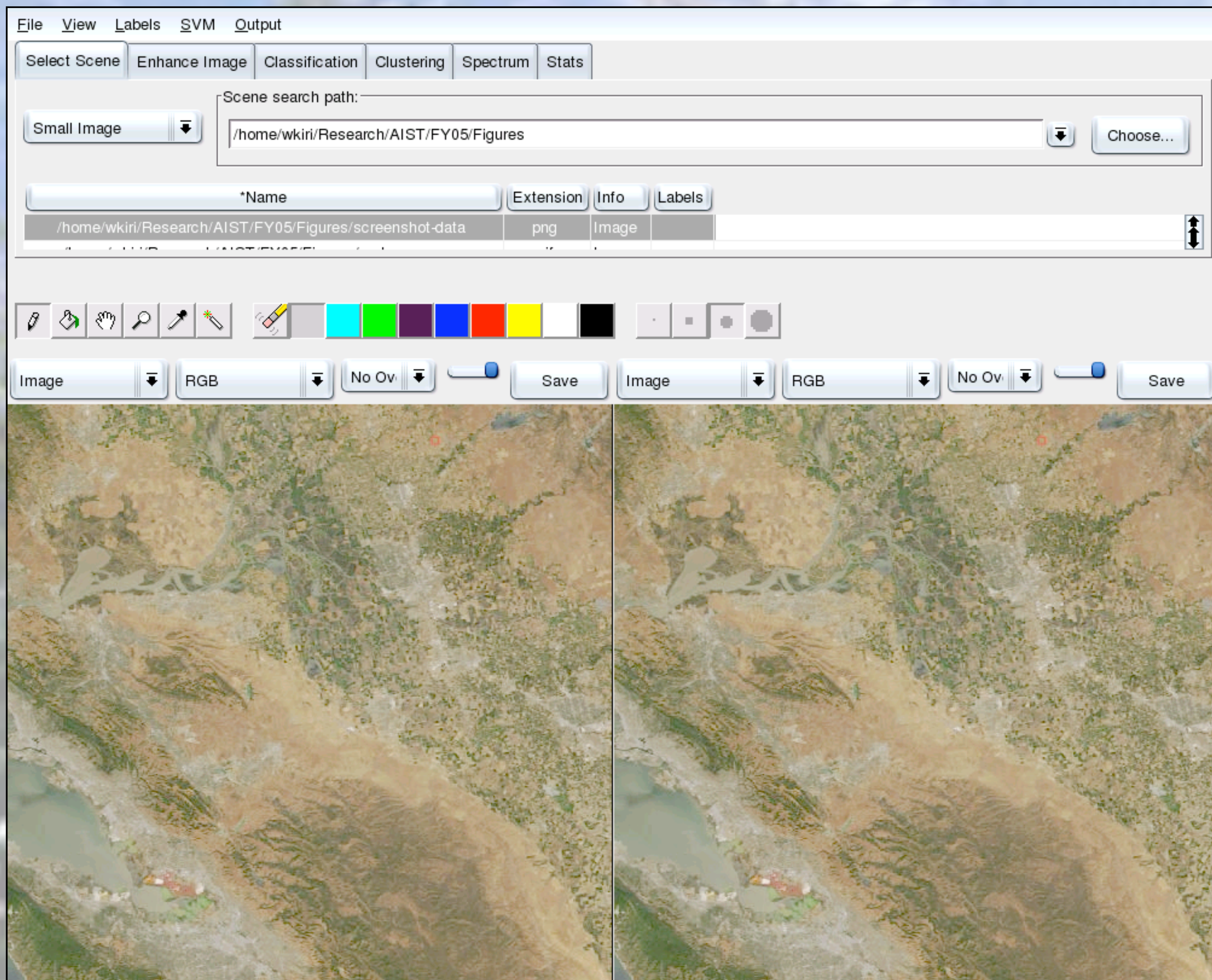
Preliminary Study: Corn Yield Prediction

- Goal: predict corn yield for 12 California counties
  - Train on data from 2002, 2003; predict for 2004
- Results:
  - Just remote sensing data (NDVI): 17.5% error rate
  - Just weather data (temperature and precipitation): 7.9%
  - Both: 7.0%



(Sorted by error rate for “Both”)

System Demo



File View Labels SVM Output
Select Scene
Enhance Image
Classification
Clustering
Spectrum
Stats
Small Image
Scene search path:
/home/wkiri/Research/AIST/FY05/Figures
Choose...
*Name
Extension
Info
Labels
/home/wkiri/Research/AIST/FY05/Figures/screenshot-data
png
Image
Image
RGB
No Ov
Save
Image
RGB
No Ov
Save

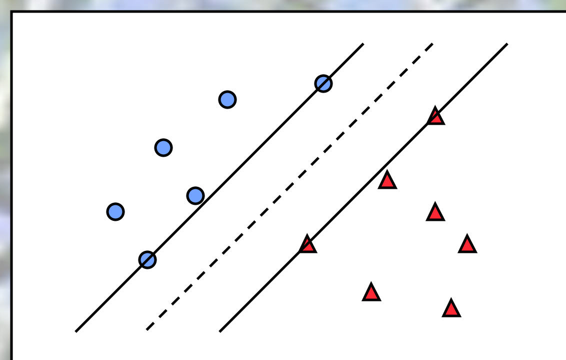
Featured Accomplishments

1. Classifier method (SVM) efficiency improvements

2. Integration of classification and clustering together
   – Example: only cluster regions known to contain vegetation

(1) SVM Efficiency Improvements

- Support Vector Machines (SVMs)
  - Identify a hyperplane that maximally separates pixels from each labeled class
  - Pixels on the boundaries are the “support vectors” (SVs)

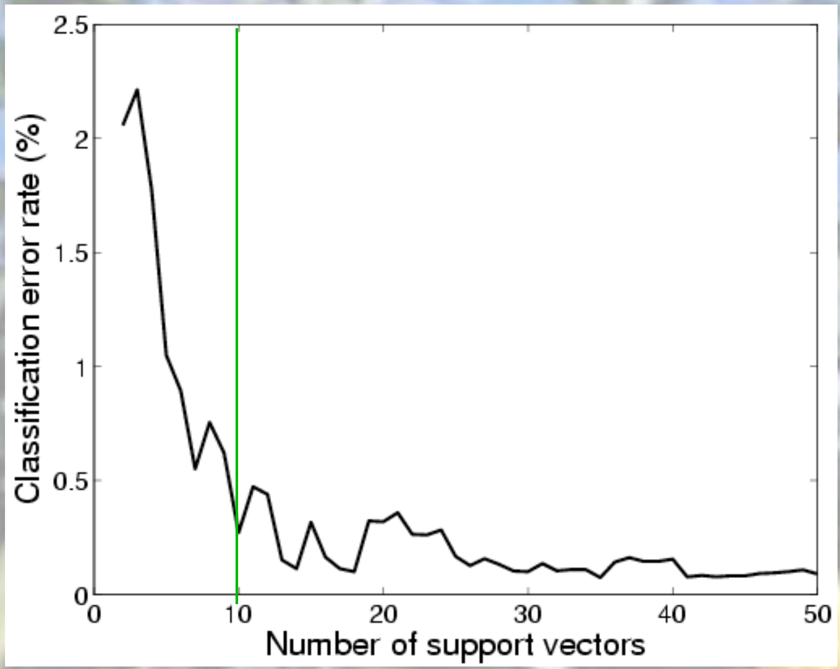


- Goal: reduce classification time (scales with number of SVs)
  - **Reduced Set** method: identify smaller set of SVs (~10x speedup, but pay pre-processing cost to find reduced set)
  - **Nearest Support Vector** method: adjust computation based on “difficulty” of item to be classified (~2x speedup)

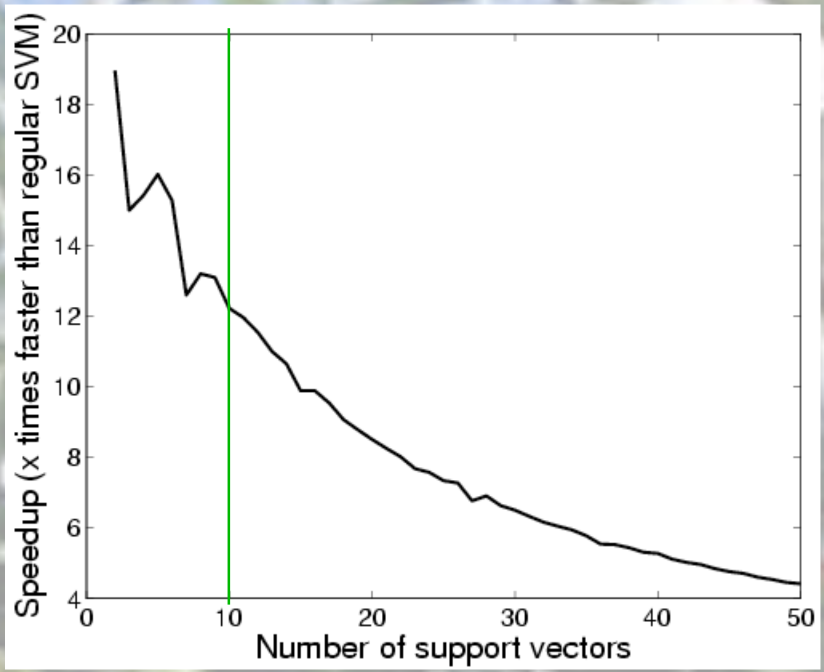
(1) SVM Efficiency Improvements: Results

- **Hybrid:** Nearest Support Vector + Reduced Set
- Tradeoff between accuracy and speed

Error rate



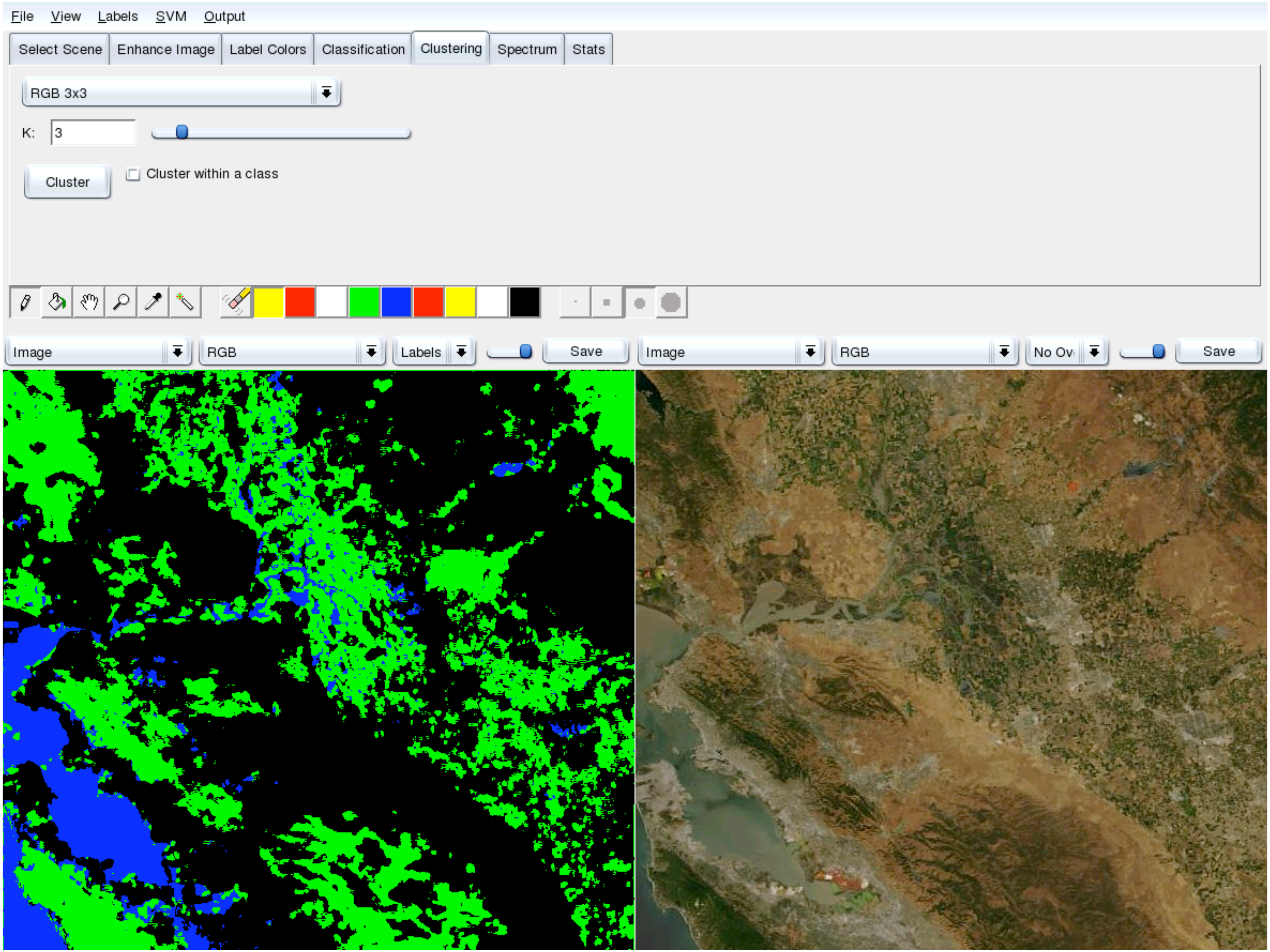
Speedup

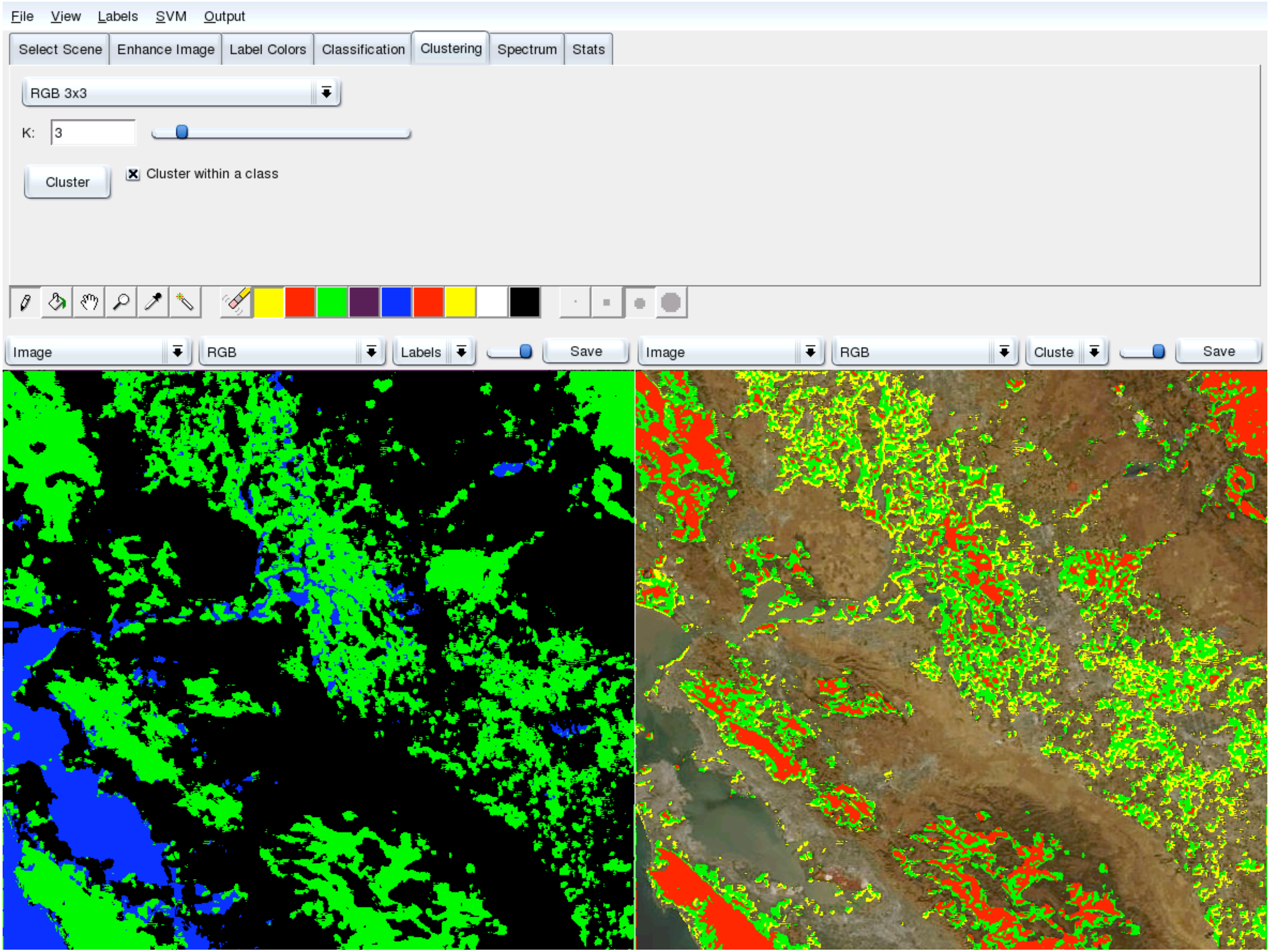


(2) Integration of classification, clustering

- Idea: analysis methods share data and results

- **Example: clustering pixels from a specific class (focus of attention)**
  - User labels and classifies the image using an SVM, then identifies a specific class to be clustered
  - Enables further exploration of class structure
    - Motivation here: identifying meaningful sub-groups within a class where we can build specialized crop prediction models
    - Clustering also returns the “average” member of each cluster, to aid in interpretation of results


File View Labels SVM Output
Select Scene
Enhance Image
Label Colors
Classification
Clustering
Spectrum
Stats
RGB 3x3
K: 3
Cluster
Cluster within a class
Image
RGB
Labels
Save
Image
RGB
No Ov
Save


File View Labels SVM Output
Select Scene
Enhance Image
Label Colors
Classification
Clustering
Spectrum
Stats
RGB 3x3
K: 3
Cluster
Cluster within a class
Image
RGB
Labels
Save
Image
RGB
Cluste
Save

Work In Progress

- Technology
  - Increase SVM, clustering **efficiency** further
    - Methods that leverage spatial relationships in the data
  - Conduct summer field study
    - Collect ground truth for different crop types in central California
    - Use this data to train a crop type classifier
    - Enables specialization of yield prediction by crop type
  - Add kriging/**interpolation** methods for to compensate for missing data or cloudy pixels
- Science
  - Analyze time series data from Kansas (105 counties) to identify **factors that impact high/low yield**
  - Integrate weather, soil properties data
    - Compare quality of predictions with and without these data sources

Long Term Goals and Benefits

- Produce a **single, integrated, graphical system** for classification, clustering, and prediction from multiple, **heterogeneous** data sources
- Demonstrate **global scalability** of enhanced SVM and clustering methods (optimized for image data)
- Demonstrate feasibility of system by **predicting crop yield** from remote sensing, weather, land cover, and soil type data bases